「パターン認識と機械学習」
2.1.1 ベータ分布

小野寺喜行

最尤推定

ベルヌーイ分布と二項分布のパラメータμの最尤推定

$$\mu_{\mathrm{ML}} = \frac{m}{N}$$

データ集合が小さいと非常に過学習してしまうので、ベイズ主義的に扱いたい

ベータ分布（beta distribution）を用いる

ベータ分布

$$\mathrm{Beta}\left(\mu \mid a,b\right)=\frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)}\mu^{a-1}\left(1-\mu\right)^{b-1}$$

次の式が満たされる

$$\int_0^1 \mathrm{Beta}\left(\mu \mid a,b\right)\mathrm{d}\mu=1$$

ベータ分布

平均

$$\mathbf{E}[\mu] = \frac{a}{a+b}$$

分散

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

ベータ分布の事後分布

ベータ分布をμの事前分布に選ぶと、μに依存する要素だけ取り出した事後分布は

$$p(\mu \mid m, l, a, b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

ただし $l = N - m$

μへの依存性が事前分布と事後分布で同じものになっている

ベータ分布の事後分布

事後分布もベータ分布として表せる

$$p(\mu \mid m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$$

これは事前分布のように振る舞える
（各観測後に、現在の事後分布を更新し、新しい事後分布を得る事が出来る）

逐次学習（sequential learning）

x の予測分布

観測データ集合Dが与えられたときの

xの予測分布

$$p(x=1 \mid D) = \int_0^1 p(x=1 \mid \mu)\, p(\mu \mid D)\, \mathrm{d}\mu$$

$$= \int_0^1 \mu\, p(\mu \mid D)\, \mathrm{d}\mu = \mathbf{E}[\mu \mid D]$$

x の予測分布

事後分布$p(\mu|D)$に対する結果と、ベータ分布の平均より

$$p\left(x=1|D\right)=\frac{m+a}{m+a+l+b}$$

$m,l\to\infty$ の極限では、最尤推定と等しくなる

ベイズ学習の特性

多くのデータを観測すればするほど事後分布が示す不確実性は恒常的に減少する

この特性は一般的か？

ベイズ学習の特性は一般的か

同時確率$p(\theta, D)$で記述されたデータ集合Dを観測したとき、パラメータθをベイズ推論する一般的な問題について考えると

$$\mathbf{E}_{\theta}\left[\theta\right] = \mathbf{E}_{D}\left[\mathbf{E}_{\theta}\left[\theta \mid D\right]\right]$$

ベイズ学習の特性は一般的か

$$\mathbf{E}_\theta\left[\theta\right]=\mathbf{E}_D\left[\mathbf{E}_\theta\left[\theta \mid D\right]\right]$$

ただし

$$\mathbf{E}_\theta\left[\theta\right]\equiv\int p\left(\theta\right)\theta\,\mathrm{d}\theta$$

$$\mathbf{E}_D\left[\mathbf{E}_\theta\left[\theta \mid D\right]\right]\equiv\int\left\{\int\theta_p\left(\theta \mid D\right)\mathrm{d}\theta\right\}p\left(D\right)\mathrm{d}D$$

θの事後平均を、データを生成する分布上で平均すると、θの事前平均に等しい事を示す

ベイズ学習の特性は一般的か

同様に、次のことも成り立つ

$$\mathrm{var}_{\theta}\left[\theta\right] = \mathrm{E}_{D}\left[\mathrm{var}_{\theta}\left[\theta \mid D\right]\right] + \mathrm{var}_{D}\left[\mathrm{E}_{\theta}\left[\theta \mid D\right]\right]$$

左辺: θ の事前分散

右辺第1項: θ の事後分散の平均

右辺第2項: θ の事後平均の分散

ベイズ学習の特性は一般的か

$$\mathrm{var}_\theta\left[\theta\right] = \mathrm{E}_D\left[\mathrm{var}_\theta\left[\theta \mid D\right]\right] + \mathrm{var}_D\left[\mathrm{E}_\theta\left[\theta \mid D\right]\right]$$

⇩

- 右辺第2項の分散は正なので、θの事後分散は左辺の事前分散より小さくなる
- 事後平均の分散が大きい程、事後分散の平均はより減少する

※ただし、この結果は平均的に成り立つだけ